To Identify What Isn't There: A Definition of Missingness Patterns and Evaluation of Missing Value Visualization

Journal Title
XX(X):1–17

DOI: 10.1177/ToBeAssigned
www.sagepub.com/
SAGE

Sara Johansson Fernstad[1]

Abstract

While missing data is a commonly occurring issue in many domains, it is a topic that has been greatly overlooked by visualization scientists. Missing data values reduce the reliability of analysis results. A range of methods exist to replace the missing values with estimated values, but their appropriateness often depend on the patterns of missingness. Increased understanding of the missingness patterns and the distribution of missing values in data may greatly improve reliability, as well as provide valuable insight into potential problems in data gathering and analyses processes, and better understanding of the data as a whole. Visualization methods have a unique possibility to support investigation and understanding of missingness patterns by making the missing values and their relationship to recorded values visible. This paper provides an overview of visualization of missing data values, and defines a set of three missingness patterns of relevance for understanding missingness in data. It also contributes a usability evaluation which compares visualization methods representing missing values and how well they help users identify missingness patterns. The results indicate differences in performance depending on the visualization method as well as missingness pattern. Recommendations for future design of missing data visualization is provided based on the outcome of the study.

Keywords

Data visualization, visual analytics, missing data, incomplete data, usability evaluation

Introduction

Missing data, meaning records that were intended to be obtained but for some reason were not, is a common issue in data analysis. It may occur in almost any domain and can cause problems such as biased results and reduced statistical rigour. The most common approaches to dealing with missing values include the removal of data records where values are missing or replacing the missing data with plausible values, commonly known as imputation. The development of statistical methods for dealing with missing data is a research topic in its own right and the selection of an appropriate method requires understanding of the mechanisms and patterns of 'missingness'. Both imputation and removal assume, to some extent, that missing values are errors that need to be dealt with. However, the fact that values are missing may in itself carry valuable information, independent of what value the record would have taken had it been recorded. The investigation of missingness patterns may provide additional understanding of complex data and gain of novel insights. In this context, visualization approaches have the potential to provide invaluable support by making the missingness and its patterns visible, and through this bring a unique potential to support decision making and knowledge generation compared to other computational approaches.

The topic of missing data has to a great extent been overlooked by the visualization society, with not much more than a handful of scientific papers presenting approaches for visual investigation of missing data over the last two decades. The importance of designing visualization systems that represent missing data has however been emphasized in recent papers by, for instance, Kandel et al.[1], Wong and Varga[2] and Fernstad and Glen[3], and was also discussed by Kirk[4]. This paper further highlights the lack of visualization approaches for analysis of missing data, providing background to the challenges involved in missing data analysis and emphasizing the benefits visual analysis may bring. Based on this, a set of patterns of particular relevance for investigation and understanding of missingness in data are defined. This is followed by a pilot study that evaluates the efficiency of three visualization approaches in context of identifying these patterns, aiming to provide guidance for future research in the area of missing data visualization. Hence, the main contributions of this paper are:

- emphasizing the lack of research and importance of missing data visualization;
- the definition of three missingness patterns (*Amount Missing*, *Joint Missingness* and *Conditional Missingness*) of relevance for understanding missingness in data;
- a usability evaluation comparing three methods for visualization of missing values.

The paper is structured as follows. The first section provides overview of missing data analysis and the challenges

[1]Digital Institute, Newcastle University, UK

Corresponding author:
Sara Johansson Fernstad, Digital Institute, School of Computing, Newcastle University, Newcastle-upon-Tyne, NE4 5TG, UK.
Email: sara.fernstad@ncl.ac.uk

involved, and presents previous research, followed by the definition of data patterns of relevance for understanding and analysis of missingness. The succeeding section presents a pilot evaluation where three visualization methods designed for representation of missing values are compared in context of the three missingness patterns. The results of the study are then analyzed and presented, followed by a discussion with guidance for future research, and conclusions in the final section.

Background and Related Work

With missing data it is unknown to the analyst what value the data would have taken if it had been observed, at best a reasonably good estimation of a plausible value can be obtained. Missing values can be caused by various reasons. In survey studies, such as demographic or consumer surveys, respondents may avoid answering particular questions or an interviewer may not ask certain questions to some of the participants. In a longitudinal study a participant may not take part in all steps, meaning all data points for a particular time point will be missing. In clinical trials subjects may drop out or be excluded from a study, for instance due to their response to treatment or if they do not follow the study protocol. In laboratory based studies physical properties may be unrecordable for certain samples, and in other situations values may not be obtainable due to processing issues or technical limitations. When analyzing data from multiple sources missingness may be caused by mismatches between databases or variations in naming conventions.

While any type of data (numerical, categorical, text, networks etc.) may contain missing records, the focus of this paper lies on missing values in multivariate (numerical) data. The visualization methods discussed as part of the evaluation are all designed for numerical data. The decision to focus on numerical data in this paper was made a) since missing values can be visualized using standard methods for categorical data, by representing them as an additional category, and b) due to the lack of available visualization methods for categorical data that specifically address missing values. Meanwhile, the missingness patterns discussed are equally applicable to numerical and categorical data.

Variables with missing values can be thought of as multi-type variables where the recorded values are of one type, such as categorical or numerical, and the missingness is a concurrent binary representation (missing or recorded). Missing values have no mean or distribution and standard statistics cannot be applied to them. Analysis of data with a large amount of missing values is a complex problem that may cause uncertainty and reliability issues. Different to many other uncertainty problems, the issues with missing data can not normally be overcome by increasing the sample size, since the number of missing values often increase alongside the data size. The process of successfully analyzing this data and properly dealing with the missing values relies on understanding the patterns and mechanisms of missingness.

The Missingness Mechanism

Different to the cause of missingness, which is the reason why data is missing in the first place, the missingness mechanism[5] is the process by which observations become missing. The missingness mechanism can be thought of as a model of how the probability of an observation being missing depends on its own value and on the values of other variables. Commonly, the mechanism of missingness is separated into three main types[5]:

- **Missing Completely at Random (MCAR)**: when the missing values occur at random and the probability of missing values depends neither on the observed nor the missing part of the data.
- **Missing at Random (MAR)**: when the probability of missingness depends on the observed data and, hence, the missingness mechanism can be expressed in terms of observed values.
- **Missing Not at Random (MNAR)**: when the probability of missingness depends on something that is not recorded. That is, if the missing values are not occurring randomly but at the same time does not depend on any part of the observed data.

MCAR situations may occur for instance when laboratory samples are dropped, whereas MAR examples may be when participants are removed from a clinical trial due to their response to a treatment. MNAR mechanisms can be particularly difficult to detect and an example of this mechanism could be when high income respondents avoid revealing their income in a demographic survey where no other questions relate to income. The missingness mechanism is rarely known prior to analysis and visual analysis of missing data may greatly facilitate understanding of the mechanism and patterns.

Identification of Missing Data

Missing values can be represented in a range of different ways in the data collection process. Depending on how the missingness information is stored, the identification of missing values can be a challenging and time consuming pre-processing step to data analysis. The missing values may be represented as empty cells or as an easily identifiable string (e.g. “N/A”, “?”) in the data table, but often they are represented by a value that may be more or less easy to identify as missing or incorrect. If unrealistic values occur in data, such as a negative product price or 25 hours of activity in a day, they can easily be spotted as incorrect, but often the missing values may be represented by a perfectly plausible value, such as zero or by simply repeating the last recorded value, making them notoriously difficult to identify. While the identification of missing values is an important challenge and difficult task, which may be greatly facilitated by visualization, it is not within the scope of this paper to analyse visualization methods for identification of missing values. The suggested missingness patterns and visualization methods used in the evaluation all assume that the missing values are already identified and explicitly marked in the data.

Dealing With Missing Data

Missing values may greatly distort the statistical properties of data, such as means and variances. Figure 1 displays examples of how missing values may affect properties,

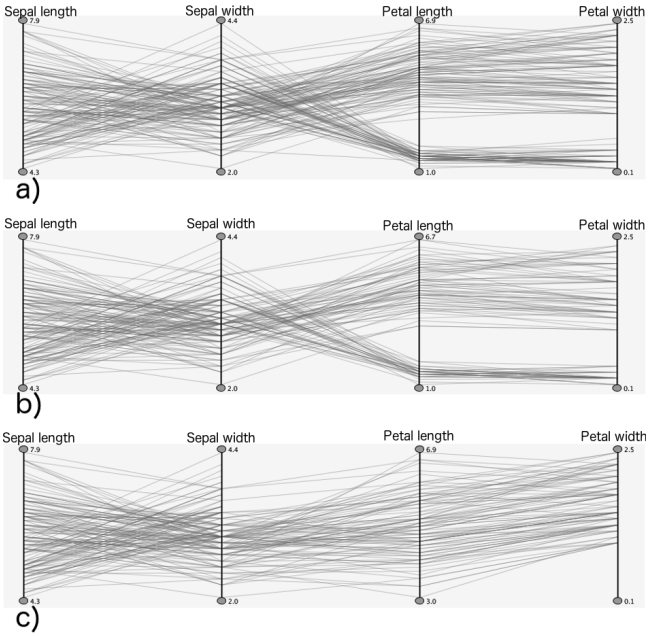


Figure 1. The effect of missing values: a) original full dataset, b) 33% missing uniformly in third variable, c) lowest 33% missing in third variable.

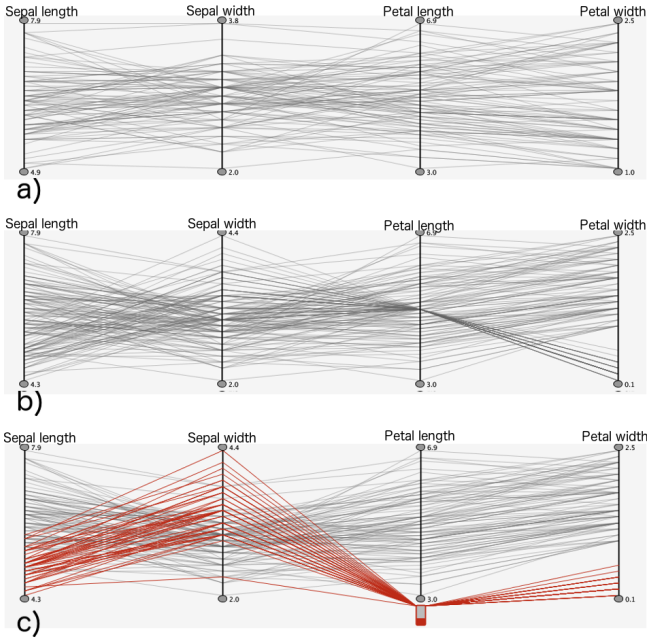


Figure 2. Examples of different ways of dealing with missing data: a) removal of all items with missing values b) imputation with mean value c) visual representation and highlighting of missing values.

using the well known iris dataset[6]. Figure 1a displays the original dataset where no values are missing; Figure 1b shows the dataset with 33% of items uniformly missing in the *PetalLength* variable (third axis from left), this is an example of MCAR; and Figure 1c displays the dataset with the lowest 33% of the *PetalLength* records missing, which can be considered an example of potential MAR since the missing values have low values for the *PetalWidth* variable (fourth axis). The mean value of the *PetalLength* variable is 3.1, 3.7 and 4.9 for the respective datasets.

The effect missingness has on analysis results depends both on the missingness mechanism and on how the missing values are handled. The degree of missingness and distribution of missing values across the dataset may also greatly affect the appropriateness of analysis methods. The two main approaches to dealing with missing values are removal and imputation. Removal, or completers analysis[5], is when data points or items that contain missing values are removed prior to analysis. Unless values are missing completely at random there is a considerable risk that removal will heavily bias the analysis results. Figure 2a displays the iris dataset with the lowest 33% of records missing in the third axis (as in Figure 1c) when all items with missing values are completely removed from the display. As visible, some of the existing patterns have become less distinguishable, such as the negative correlation between the first and second axis.

Imputation, on the other hand, is when missing values are replaced by estimates. There are numerous methods available for imputation, ranging from simple replacement with arithmetic mean or random draws from representative distributions, to complex multiple imputation methods where several imputed versions of the dataset are combined following a set of rules[5]. Depending on which imputation model is used, the imputed values may bias analysis results and the variability in the imputed dataset may often be too small. This is a particular issue with simple approaches, such as mean imputation, and when data is not missing completely at random. Figure 2b displays an example of imputation using the same dataset as in Figures 1c and 2a, in figure 2b missing values are replaced by the mean value of the variable, creating an artificial cluster around the mean. It is clearly visible from Figures 2a and 2b that the selected method for dealing with missing values can have a big impact on analysis results.

Visualization of Missing Data

The fact that values are missing may provide valuable information, such as highlighting potential problems in data gathering, pre-processing and analysis processes. Fielding et al.[7] highlight that the absence of data can be informative, particularly in health-related questionnaire studies where participants that are unwell may be less likely to respond. Another example was provided by Djurcilov and Pang[8] who discuss visualization of meteorological datasets where a missing value was an indication that no phenomena were observable. In this context they suggested that missing values should be presented such that the user is alerted about the missing data, rather than estimating a value. Some recent texts[1–4] also highlight the need of visualization that enable exploration and further understanding of missingness in data, but only a small number of publications to date describe methods for missing data visualization.

Twiddy et al.[9] were among the first to address challenges related to visualization of missing data. They aimed to avoid the risk of misinterpretation caused by replacing missing data with interpolated values, and presented an approach for missing data visualization where recorded and missing values were visually separated using a colourscheme where the pop-out effect of the missingness representation was reduced. The MANET software[10,11]

was another early example that aimed to make the user aware of the incompleteness of data by incorporating visual representations of missing values in, for example, bar charts and scatter plots. An approach partially related to MANET was later presented by Templ, Alfons and Filzmoser[12]. Their R-package VIM (Visualization and Imputation of Missing values) was designed for exploration of missingness structures and utilizes various visual attributes to highlight missingness in common visual representations such as histograms, scatter plots and parallel coordinates. The package miP (multiple imputation plots)[13] visualize imputation results from a range of packages using VIM. Schulz et al.[14] defined missing data as one of several data descriptors, and showed how missing values could be presented in parallel coordinates using a method similar to VIM. Additionally, some R-packages for imputation of missing values include graphical user interfaces for manipulation and control of imputation methods, including migui[15], that provides an interface for the mi package[16]; and AmeliaView, which is a function in the Amelia package[17]. While the user interfaces are helpful in facilitating imputation, they are not directly supporting exploration and identification of missingness patterns.

A slightly different approach was taken in the xGobi system[18], later followed by gGobi[19], where missing data was represented by imputed values while a separate linked view was used to keep track of the location and existence of missing data, utilising interactive features such as brushing, zooming and labelling to support exploration of missingness. Building upon this, Cheng et al.[20] developed the R-package MissingDataGUI, which supports exploration of missing data structures using summaries and static graphics where missing values are imputed and distinguished from recorded values by colour. Cedilnik and Rheingans[21] used procedurally generated annotations to represent uncertainty information, and represented missing data with a distance based probability value. A similar approach was taken by Xie et al.[22] who introduced a visualization system focusing on data quality, including missingness and uncertainty, and obtained quality values for missing data using imputation methods. Arbesser et al.[23] presented a linked views system for assessment of data quality, where missing data is one of a number of quality classes that are distinguished through colour. Wang and Wang[24] addressed visualization of missingness in classification data. Their main focus lay in identification of whether the missing values are randomly distributed, unevenly distributed across variables or biased towards a particular class. They address the high dimensionality of classification data by utilizing self-organizing maps (SOM)[25] to cluster a missingness representation of the data.

Another aspect of missing data that is highly relevant in visualization is the effect the representation of missing values has on the interpretation of data. This was discussed by Eaton et al.[26] who defined three levels of impact that missing data may have on visualization: 1) when the missingness is perceivable from the visual representation, 2) when nothing indicates the existence of missing values and 3) when the missingness propagates to other items and affects their visual appearance. Figure 3 provides examples of these three levels. Eaton et al. also present a user study where they compare

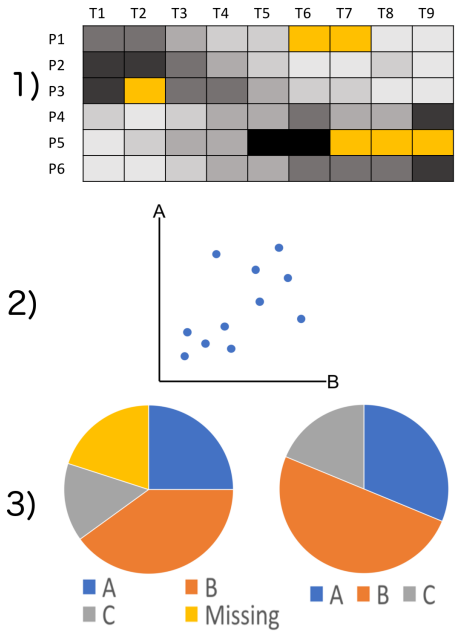


Figure 3. Levels of impact of missing values on visualization: 1) missingness is perceivable in the visualization: missing values are represented in yellow in the heat map; 2) nothing indicates the existence of missing values: data points with missing values are not drawn; and 3) missingness propagates to other items and affect their appearance: the left pie chart displays missing data as a separate category, whereas the missing data has been removed in the right pie chart, which affects the relative size of the other three categories.

three different approaches to representing missing values in line graphs: *misleading* where the missing values are replaced by zero, *absent* where missing values are omitted from the visual representation, and *coded* where missing values are omitted and the next valid point has a mark and provides information why the prior points are missing. Their results imply that poor indication of missingness in a visual display have a clearly negative effect on interpretation and that the user may not realize that a value is missing if it is replaced by a value. Based on this, Eaton et al. suggested that visual representations should be enhanced either by dedicated visual attributes (using colour, shape etc. in a visual representation), annotation (using text or graphic information outside of the visual representation) or animation to indicate the existence of missing data, but provide no recommendations as to what type of visual attribute may be most appropriate to support understanding of missingness.

Missingness Patterns in Literature

The missingness mechanisms highlight structures of potential relevance for deciding how to deal with missing data and may be useful for understanding the effect of imputation and deletion. The mechanisms are, however, fairly complex and rarely known prior to analysis, and may not be directly applicable to an exploratory analysis approach. While the MCAR and MAR situations may possibly be identified by analyzing the dataset and the missingness, MNAR situations are inevitably difficult to identify by analyzing the data since their explanation is not available within the dataset.

Exploratory analysis aims to identify and summarize data patterns to support hypothesis generation, and a different set of patterns may be more relevant in the context of exploring and analyzing missingness.

Wang and Wang[24] highlighted the relevance of the distribution of missing values within the dataset, particularly focusing on if it is randomly or unevenly distributed across variables and classes. They suggested three missingness patterns for analysis of missing values in the context of classification data: 1) *Missing At Random*, when values are randomly distributed in the sample space (note that this is not the same as the MAR missingness mechanism), 2) *Uneven Symmetric Missing*, when values are missing more often in some variables and missing values may be correlated across variables, and 3) *Uneven Asymmetric Missing*, when values are missing unevenly in the data and may be biased towards a particular class. In another paper, Wang and Wang[27] discuss the impact of missing values on data mining tasks and the value of knowledge about the patterns of missingness and their potential impact on results. They advocate a problem-driven approach and list four concepts of relevance for understanding the impact of missing values: 1) *Reliability* addresses the scope of the missing values in context of the problem, where the problem is defined only based on recorded values. It can, as such, be thought of as a ratio between missing and recorded in context of the problem domain, and can for instance include comparison between the number of missing values in a variable and the number of records that are used to identify the problem. 2) *Hiding*, which is a concept to reveal how likely it is that a data item, within a certain range of one variable, has a missing value in another variable. 3) *Complementing* is a concept aimed to reveal which variables are most likely to have missing values at the same time. 4) *Conditional Effects* are the potential changes caused by the missing values to the understanding of the problem, which can be examined by replacing missing values with different possible values and observe how it changes the problem. Theus et al.[11] point out that in a pairwise relationship involving missing data, each data point can belong to one of four different states: 1) both values are recorded (not missing value), 2) the x-value was recorded but not the y-value, 3) the y-value was recorded but not the x-value, and 4) neither of the values were recorded. In state 2, 3 and 4 we may be able to draw conclusions in relation to missingness patterns based on the data; particularly if there are relationships between missing values and recorded values and if there are relationships between missing in one variable and missing in another variable.

Three main conclusions can be drawn based on the prior research. Firstly, the distribution of missing values is important for missing data analysis, both in terms of even or uneven distribution, as well as the ratio between missing and recorded values. Understanding of distribution may support insight into whether values are missing at random, as well as into the reliability of conclusions drawn from the recorded data. For instance, the awareness of unevenly distributed missing values and a high ratio of missing values in certain variables, may reduce the reliability of results driven by recorded values in those variables. Secondly, the relationship between missing values in one variable and the values of recorded data in another variable is suggested as relevant. This has similarities to correlation patterns and can, to some extent, be thought of as the correlation between missing and recorded. Common correlation metrics can, however, not be directly applied to describe the pattern, since it concerns the relationship between a binary value and a numerical/categorical value range. Understanding of such missing–recorded relationships can help explain why values are missing (i.e. study participants with strong reactions to a treatment may stop the treatment, leading to missing values for those participants) and guide the selection of appropriate imputation method (i.e. if missing in A tend to have low recorded values in B, the imputation result may be better if estimated based on items with low values in B, rather than on items with both low and high values in B). Thirdly, the co-occurrence of missing values in multiple variables can support understanding of multivariate missingness patterns and identification of clusters of items where missingness is an issue, which may for instance support both identification of issues in the data gathering process (i.e. when errors in one step of a measurement process propagates to other measurements) and guide the selection of appropriate method for dealing with the missingness (i.e. a group of items with co-occurring missing values may need to be dealt with differently than items where co-occurrence is less common).

Patterns of Missingness

Common methods used to understand general structures and relationships in data may not be appropriate as descriptors for missingness patterns, due to the dual nature of incomplete data where variables may concurrently hold both binary values (missing or recorded) and, for instance, numerical values. The missingness patterns dicsussed in the previous section have similarities with common data patterns used to describe complete data (distribution, coinciding values and correlation). They are, however, not identical as they are designed and used in the context of single-typed data and not designed to deal with and support analysis of missing values and would, hence, require modification to be usable. None of the missingness patterns discussed in the previous section completely cover distribution, co-occurring missing and missing–recorded relationships in a satisfactory way for multivariate data analysis.

The missingness mechanisms are difficult to translate to data analysis tasks and hard to identify; Wang and Wang's[24] missingness patterns for classified data only covers distribution patterns, and are specifically design for classification tasks; their second paper[27] cover all three patterns but are discussed mainly in the context of understanding the impact of missing values and are defined as part of a two-step analysis process, thus not directly applicable to generic data analysis tasks; and the description made by Theus et al.[11] does not address the distribution of missing values. Based on this and on the conclusions in the previous section, this paper contributes the definition of three novel missingness patterns (*Amount Missing*, *Joint Missingness* and *Conditional Missingness*), as briefly introduced by Fernstad and Glen[3]. The relevance of these patterns for describing common missingness structures was confirmed through informal interviews with data science

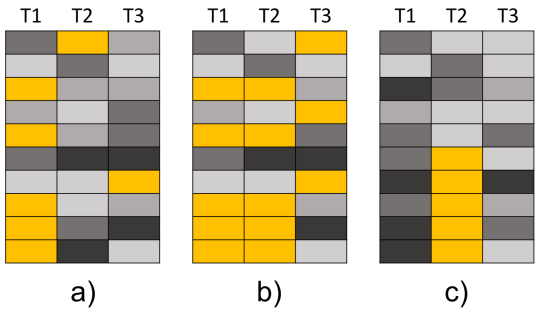


Figure 4. Examples of the three missingness patterns (yellow cells corresponding to missing values): a) Amount Missing: 50% of values are missing in T1 (left), while only 10% are missing in T2 (centre) and T3 (right). b) Joint Missingness: There is a pattern of joint missingness between T1 (left) and T2 (centre), but not between T2 and T3 (right). c) Conditional Missingness: The missing values in T2 (centre) are conditional upon high values in T1 (left, dark cells), but not upon values in T3 (right, mix of light and dark cells).

practitioners who deal with missing data as part of their everyday work. This section will further extend upon these patterns and put them in context of previous literature.

Amount Missing

Amount Missing (AM) refers to the relative number of data values that are missing, either in a data item or in a variable. Insight into the amount of missing data provides understanding of the distribution of missing values within the dataset and can be useful for providing an indication of a potential MCAR situation (or exclusion of the possibility of MCAR), since the distribution of missing values across the dataset would be even if they are missing completely at random, while unevenly distributed missingness would outrule MCAR. Wang and Wang[24,27] highlights the relevance of missingness distribution, both in context of classification and through their reliability concept.

Insight into the amount missing values in variables provides understanding of the distribution of missing values across the variables of the dataset and can, as such, be useful for identification of variables where the missingness may be particularly important or difficult to deal with due to a large amount of missing values. Similarly, it can highlight subsets of data where conclusions drawn from the recorded data may be particularly uncertain, due to being based on a relatively small amount of recorded values. When applied to data items the AM pattern may also highlight individual samples and outliers with high relative amounts of missing data that may be too incomplete to include in the analysis and, hence, may be better to remove. Figure 4a displays an example of AM in variables where the missingness is not evenly distributed, with 50% missing (yellow cells) in T1 while only 10% missing in T2 and T3.

Joint Missingness

The second suggested pattern of interest for analysis of missingness in data is *Joint Missingness* (JM). JM is a pairwise or multivariate pattern that refers to the amount of data items that have concurrently missing values in more than one variable at the same time. Figure 4b displays an example of joint missingness between variables T1 and T2, whereas the missingness between T2 and T3 is not joint (yellow cells corresponding to missing values). To some extent, JM is related to the correlation or association between categorical variables, and coinciding features as used in text analysis for example, but with the difference that only one particular value (missing) is of interest and not any coinciding values, and with the missingness being of a different data type than the recorded values of the variable. JM patterns would indicate that data is not missing completely at random and that there may be a potential MNAR situation where the missingness depends on something that is not recorded in the data. It is closely related to the concept of complementing[27] and corresponds to the fourth state of pairwise missingness relationships described by Theus et al.[11], where neither the x-value nor the y-value is recorded.

Insight into the JM in a dataset helps understanding multivariate missingness patterns; which may, for instance, be when participants who refuse to answer a particular question in a questionnaire (such as their income) also refuse to answer other particular questions (such as the value of their house). As previously mentioned, it can also support identification of issues in the data gathering process that cause missingness to propagate across the data and facilitate identification of subsets of data where the missingness may need to be dealt with differently from subsets with less joint missingness.

Conditional Missingness

Like JM, *Conditional Missingness* (CM) is a pairwise or multivariate pattern. It refers to the relationship between items that are missing in one variable and the recorded value of those items in another variable. The CM pattern is relevant for understanding relationships between missing and recorded values and aims to describe patterns where the probability of missingness is conditional, or dependent, upon the recorded values of another variable. As such it has similarities with correlation patterns, but is different in terms of requiring one type of data for one variable (missing) and another type of data for the other (numerical, categorical etc.), as a consequence, standard correlation metrics cannot be utilised to identify the pattern. Furthermore, correlation of missing values may equally refer to joint missingness and, hence, the correlation concept does not distinguish the two patterns clearly enough. Figure 4c shows an example where the missing values in T2 (yellow cells) are conditional upon high values in T1 (dark cells), while they are not conditional upon values in T3 (mix of light and dark cells). CM relates both to the MAR mechanism, where missing data depends on recorded data, and to Wang and Wang's concept of hiding[27]. It is also described as the second and third state of missingness in pairs of variables by Theus et al.[11].

Understanding of CM patterns can be useful for deciding how to deal with the missingness in terms of imputation or deletion, as well as for understanding the cause of missingness. For example, if participants with high income refuse to state their income in a questionnaire, but provide detail about their property tax, we may potentially see a relationship between missing income and high property tax. By investigating the recorded income values for participants with high property tax we may be able to estimate a more

reliable value for the missing values than if the estimation is based on all recorded values.

Evaluation of Missing Value Visualization

Eaton et al.[26] conducted a study to evaluate the three levels of impact that missing data can have on visualization methods (*misleading*, *absent* or *coded*), using line graphs as their basic visual representation. Based on their results they suggest that visual representations should be enhanced by dedicated visual attributes, annotation or animation to indicate the existence of missing data. Of these three, the use of dedicated visual attributes is by far the most common. The study presented here aims to evaluate the effectiveness, in terms of clear representation of relevant patterns, of three visualization methods where different visual attributes are used to represent missing values.

Visualization Methods

One of the more recent and most mature tools for visualization of missing data is the R-package VIM[28]. VIM provides a range of visualization methods (such as scatter plots, parallel coordinates, matrix plots) that are enhanced with visual attributes to represent the missing values in conjunction with the recorded data. The basic visual attributes used, on their own or in combination, to represent or emphasize missing values include:

- *Location* - the positioning of missing values at a dedicated location in the visualization, separating it from the recorded values. An immediate risk with this is if location has a meaning for the recorded values (i.e. the further to the right the higher the recorded value) and, hence, the location of the missing value may erroneously indicate a high or low value, rather than a missing value. The misleading visualization as defined in Eaton et al. use location to indicate missingness, but more commonly the location would be outside of the normal value range of the visualization.
- *Colour* - the use of a specific colour to highlight that values are missing. Depending on what colour scheme is used, a potential issue may be that the colour of missing values may distract attention from the overall data patterns, as highlighted by Twiddy et al.[9]. Colour is usually not considered an appropriate attribute for representation of exact numerical values, but can be useful for representation of a smaller number of (unordered) categories and, as a salient pre-attentive feature, effective for highlighting values of interest.
- *Size* - size is commonly used either for representing a numerical value or a frequency; in context of missing values it is mainly meaningful as frequency representation since missing values have no numerical value. For visualization methods where data items with the same values are drawn separately, rather than on top of each other, a sense of size or frequency will be created since the number of pixels used increase with the number of items.
- *Connection* - multiple data values may be linked in a visualization, indicating connectedness. While this is not normally a visual attribute used to represent missing values as such, it is commonly used for detection of multivariate patterns and may, thus, be relevant for identification of multivariate missingness patterns.

Table 1. Summary of Attributes of Visualization Methods

	Marginplot Matrix	Matrix Plot	Parallel Coordinates	Relevant for pattern
Colour represent missing	partly	yes	partly	All
Location represent missing	yes	no	yes	All
Size represent missing	partly	partly	no	AM, JM
Items with same value	on top	separate	on top	AM, JM
Connect across variables	no	partly	yes	JM, CM
Connect missing & recorded	partly	partly	yes	CM
Separate missing & recorded	yes	no	partly	CM

Three visualization methods from VIM were selected for the evaluation, all being based on common visualization methods for multivariate data analysis that have been enhanced with different visual attributes to represent missing values. The methods and their approaches to representing missing values are described in the following sections, with main visual attributes and features summarized in Table 1.

Marginplot Matrix: The Marginplot Matrix is similar to a scatter plot matrix where the individual scatter plots are enhanced with visual attributes representing missing values. Figure 5 displays an example of how the missing values are represented in a single marginplot. The representation utilize a categorization similar to the one suggested by Theus et al.[11]. Items with values that are recorded in both variables of the scatter plot are represented by blue points in the main body of the plot, while items with missing values are represented by red points in the margins. Items with missing values in one variable but not the other are represented by red points along the margin for which they have recorded values, positioned according to the recorded value. Hence, items with missing y-values but recorded x-values are positioned in the bottom margin, along the x-axis. From figure 5 we can tell that the majority of items with missing y-values have relatively high x-values (represented by the red points in the right half of the bottom margin) while a smaller number have low x-values (represented by the red points in the left half of the bottom margin). The plot also includes additional

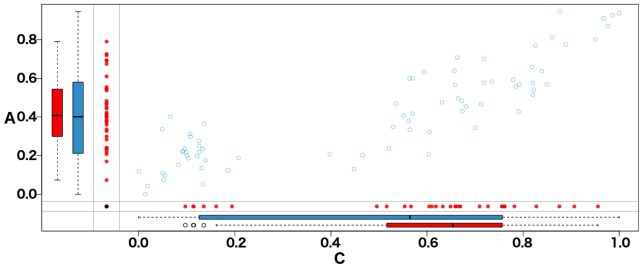


Figure 5. The marginplot is similar to a scatter plot and displays items with missing values in red. Box-plots are used to display the distribution of recorded values (blue) and missing values (red). The representations in the bottom margin correspond to items that are recorded for the x-axis variable, and representations in the left hand margin correspond to items that are recorded for the y-axis variable.

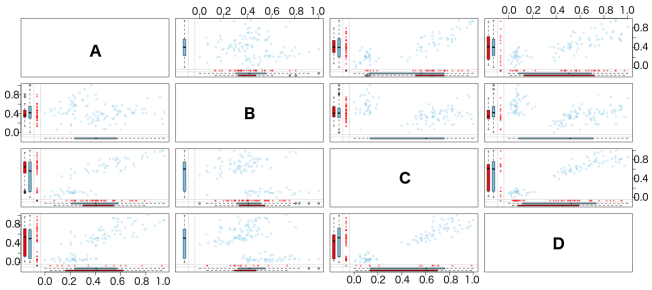


Figure 6. A Marginplot Matrix displaying a dataset with four variables, including marginplots for all pairs of variables.

margin box-plots that display the distribution of values for items recorded in both variables (blue box-plots) and items recorded in only one variable (red box-plots). The dark red point in the bottom left corner of the margin represent items with missing values in both variables.

The Marginplot Matrix (figure 6) is essentially a matrix of marginplots, where all plots in the same row have the same x-axis variable and all plots in a column have the same y-axis variable. The plot indicates missing values using a combination of colour and location, where the missing data is partly separate from the recorded data rather than mixed with the recorded items, this could possibly make it less effective for identifying relationships between missing and recorded as the user will need to switch context. Additionally, the small multiples approach has a general drawback in terms of the size of the plots as the number of variables increase, as well as diagonal matrix layouts including duplicates of all plots. On the other hand, since all pairwise relationships are represented, patterns related to variable pairs may be easier to investigate compared to, for instance, using parallel coordinates where relationships between non-adjacent axes may be less easy to perceive. Size is used as a visual attribute to represent the distribution of values that are recorded in at least one variable, through the box-plots. Items with missing values in both variables are represented at a single position. This may potentially make it difficult to estimate the number of items that are missing in both variables of a variable pair, having to take the total number of visible data points (recorded and part recorded) in the plots into account to estimate how many are likely missing in both variables.

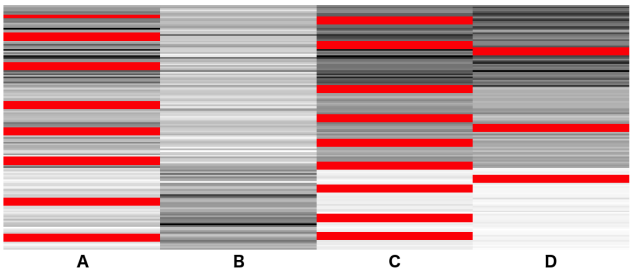


Figure 7. A Matrix Plot displaying a dataset with four variables. Recorded values are represented by grey scale, light grey corresponding to low values and dark grey corresponding to high values. Missing values are represented by red.

Matrix Plot: The Matrix Plot is a heatmap, or tabular plot, were columns represent variables and rows represent data items. The colour of a cell represents the value of an item for corresponding variable. Recorded values are represented by grey scale, with dark grey corresponding to high values and light grey corresponding to low values. Missing values are represented by red colour, as shown in figure 7. Differently to the Marginplot Matrix, the missing values are represented only by colour and are mixed with the recorded values in the Matrix Plot. Due to this, it may potentially be easier to perceive patterns relating to relationships between missing and recorded values in the Matrix Plot compared to the Marginplot Matrix. Furthermore, the values are never drawn on top of each other in the Matrix Plot, hence avoiding the issue occurring in Marginplot Matrix when values are missing for both variables as well as generating an impression of size (or number of pixels) relative to the number of missing values, which may be beneficial for frequency related tasks. The interpretability of pairwise patterns may, however, be more of an issue in the Matrix Plot when it comes to pairs of variable that are not adjacent in the display.

Parallel Coordinates: The Parallel Coordinates in VIM represent missing values through a combination of location and colour. Missing values are located above and separate from the variable axis for which it is missing. Additionally, the line colouring is based on whether a value is missing or recorded in a variable of choice, red corresponding to missing values and blue corresponding to recorded values. Figure 8 displays an example where colouring is based on whether values are missing or recorded in variable C (third axis from left). In the figure it is visible that variables A, C and D (first, third and fourth axis) have missing values, since they all have lines intersecting above the axes, while there are no missing values for variable B (second axis from left). The representation of missing values in Parallel Coordinates displays missing values more or less mixed with the recorded values, which may facilitate identification of patterns relating to a combination of missing and recorded values. Similarly to the Matrix Plot, a known issue with Parallel Coordinates is the identification of patterns relating to pairs of variables when the variables are not adjacent, but values are on the other hand connected across the axes which may facilitate identification of multivariate patterns. An issue when using location to represent missing values, as previously discussed, is the risk of misinterpretation caused by the location; by

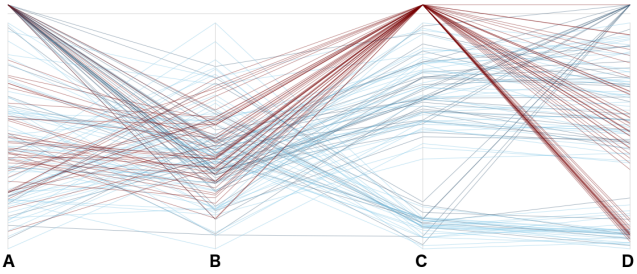


Figure 8. A Parallel Coordinates plot displaying a dataset with four variables. Missing values are represented above the axis for which it is missing. Line colouring is based on if the item has recorded or missing value for a selected variable. In this figure, colouring is based in the third variable from the left (C), meaning that items that have missing values in C are red, while items with recorded values in C are blue.

representing missing values above the axis (particularly when the boundary between recorded and missing is fairly vague) the value may easily be misinterpreted as a high value rather than a missing value. Furthermore, since missing values are represented at a single intersection point above the axis, it may be difficult to perceive the number of values that are missing for a single variable. The user will need to take the distribution of lines in adjacent axes into account to estimate the number of missing values; this will potentially also obstruct the interpretation of frequency in joint missingness patterns.

Motivation: The choice of these three visualization methods was made based on them being common visualization methods that are able to display the dataset directly as it is, without utilizing summaries or aggregation, and as they can be used for general data analysis tasks as well as missing data analysis, thus imitating a generic data analysis situation. Other visualization methods could have been used, such as aggregated bar charts and tables displaying only information about the missing values, but due to their lack of support for generic data analysis (including both recorded and missing values) and building upon the idea of enhancing visualization with visual attributes, the three selected were considered better candidates for the evaluation.

In addition to being able to support generic data analysis the ability to display recorded as well as missing values, in contrast to displaying only missing values, has some benefits for missing data analysis. As previously discussed, there may exist relationships between missing and recorded values (as in an MAR situation) where the recorded values are just as important as the missing values for identifying the pattern and for dealing with the missing values. Furthermore, it can support the identification of missing values and the evaluation of the reliability of previous missingness identification. The identification of missing values can be difficult, as discussed in the *Identification of Missing Data* section, and it is not unlikely that an incomplete dataset may include values that erroneously have been classified as recorded instead of missing. The visualization of recorded values can facilitate the identification of such potentially misclassified data. An example of this is shown in figure 9 where data from a study of skill learning in computer game playing is displayed, including a number of measures extracted from data from 3360 players at different levels of expertise[29]. The visualization used in this example is a parallel coordinates plot where missing values are represented below the axis, and items with missing values in the *Total Hours* variable (sixth from left) are highlighted in purple across the dataset. An interesting pattern (highlighted with yellow) can be seen in the *Game ID* and *League Index* variables, with almost all items that have missing values for Total Hours being separated from the rest of the data, with Game ID above 10000 (other items having a continuous spread of values below 9270) and a League Index of 8 (the number of leagues reported in the paper being 7). While it is not completely clear if these values are in fact missing values that have been represented by a numeric value as part of the data collection or pre-processing, it indicates an uncertainty or data processing issue that should be investigated further and that wouldn't have been identified if only displaying missing values.

The Study

The study presented in this section is a pilot study aiming to evaluate the ability of visualization methods to clearly display patterns of relevance for missing data analysis. With the intent to support and guide further research into visualization of missing data. The evaluation does not aim to exhaustively examine all aspects of usability for the visualization methods, but rather focus on whether the patterns of interest can be correctly identified. More specifically, the study evaluates the performance of three visualization methods (Marginplot Matrix, Matrix Plot, Parallel Coordinates) when carrying out tasks related to the three missingness patterns (*Amount Missing* (AM), *Joint Missingness* (JM), *Conditional Missingness* (CM)), as defined in the *Patterns of Missingness* section of this paper. Ethical approval was received prior to carrying out the study.

Hypotheses: Based on the visual attributes of the three visualization methods to be evaluated (summarized in Table 1), and the variation in information required to identify the three missingness patterns, the main hypothesis underlying the study is that there will be a difference in performance for the visualization methods, and that this difference is dependent on which pattern is being examined, as discussed below.

For AM the frequency of missing data in a variable has to be identified. This requires understanding of the number of missing values in relation to the number of recorded values in the variable, and the ability to do this is impacted by how items that have the same value are represented. Both Marginplot Matrix and Parallel Coordinates draw items with the same value on top of each other while Matrix Plot draws them separately, where the number of pixels used to draw them increase with the number of values. However, for items with missing in one variable but not the other, Marginplot Matrix does not draw the missing values on top of each other. Based on this it is believed that the Matrix Plot will perform better than Parallel Coordinates and Marginplot Matrix for AM tasks, while the Marginplot Matrix will perform better than Parallel Coordinates.

JM tasks involves the identification of two variables that have a large number of jointly missing values, which requires

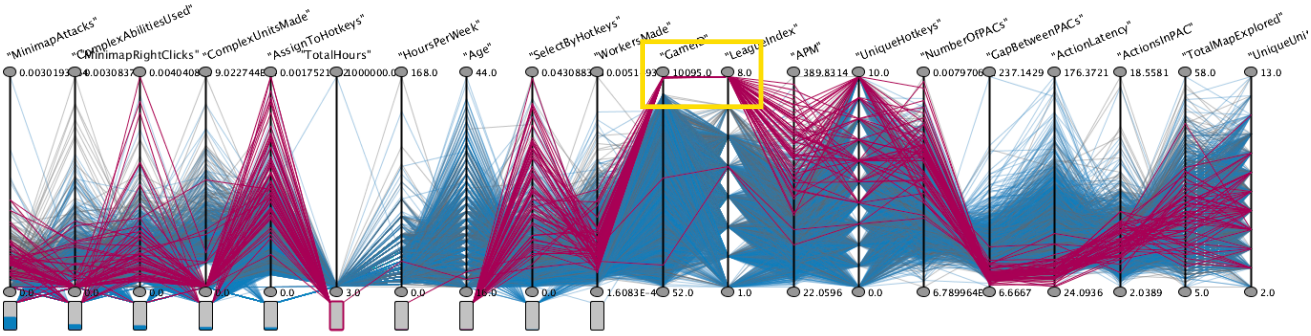


Figure 9. Parallel coordinates displaying data from a computer game skill learning study[29]. Items with missing values in the *Total Hours* (sixth variable) are highlighted in purple, an interesting pattern relating to the missing values is highlighted in the yellow box.

both understanding of the frequency of missing values and the ability to link missing values across multiple variables. Matrix Plots have the ability to display frequency, while Parallel Coordinates have stronger visual connections across variables. Based on this and since Matrix Plot still has some ability to connect across variables, it is believed that Matrix Plot will perform best for JM tasks, and that Parallel Coordinates will perform better than Marginplot Matrix.

When it comes to CM patterns the interest lies in identifying a data trend, in terms of recorded values, that may explain why values are missing in a particular variable. This requires the ability to connect missing and recorded values across multiple variables, and a hypothesis based on this is that Parallel Coordinates will perform best for CM tasks. It is also expected that the box-plots in Marginplot Matrix will be of some benefit for linking missing and recorded values and, hence, that Marginplot Matrix will perform better than Matrix Plot for CM tasks.

Tasks and Data: The participants of the study were to carry out tasks related to identifying or understanding the three missingness patterns using the three visualization methods; hence questions relating to the *Amount Missing*, *Joint Missingness* and *Conditional Missingness* were asked. The following three base questions were used for this purpose:

- **AM:** Approximately how much data is missing in attribute X?
- **JM:** With which attribute does attribute X have the most jointly missing values?
- **CM:** Which trend in which attribute is most likely to be accountable for the missing data in attribute X?

Answers to the questions were designed as multiple choice where one answer was accurate, and the rest range from potentially accurate to definitely wrong.

To ensure that the missingness patterns in the data used for the study were identifiable, while maintaining realistic patterns in the recorded part of the data, publicly available datasets were used and missingness was generated through controlled removal of values. It was expected that the scalability would mainly be dependent on limitations in the basic visualization methods, rather than on how the missing values are represented, since the majority of screen space is used similarly to standard scatterplot matrices, heatmaps and parallel coordinates. Fairly small datasets were used, for which visual clutter should not be an issue, aiming to minimize the impact of scalability limitations caused by the basic visualization methods. A total of 36 datasets were generated, 18 based on the *Iris* dataset, which includes 4 variables and 150 items, and 18 based on the training part of the *User Knowledge Modelling* (UKM) dataset, which includes 5 variables and 258 items. Varying levels of uniformly distributed noise, with a noise level between 1% and 15%, was randomly added to the data. The datasets were separated into three groups, one for each missingness pattern, with 6 Iris based datasets and 6 UKM based datasets in each group. Missingness patterns were then created by replacing numerical values with a "NA" string, with between 0% and 40% of values removed from each variable to generate the missingness patterns. Prior to removal of values, the missingness structures of the variables in all datasets were defined (included as supplementary material). For AM patterns the relative amount missing for each variable was first defined, and then converted into the actual number of values to be removed. For JM the intial step was also to define the relative amount missing for each variable, followed by the relative amount of jointly missing values for each variable pair and then the corresponding number of values to be removed for variables and pairs. For CM only one of the variables had missing values, since the question was related only to missing in one particular variable, and the first step was to define the relative amount missing for this variable, which was then converted into an actual number of values to remove. The final step for CM was to define the relationship between missing values and recorded values in the other variables, these were recorded either as high, low or none. The data was then matched with the visualization methods such that each plot was assigned two Iris and two UKM datasets from each pattern group. The variable names of the original datasets were anonymized and replaced with letters to avoid impact of preconceptions based on variable names.

Experimental Design: The study was designed as a 2-factor within-subject design, where the two factors were missingness pattern (AM, JM, CM) and visualization method (Marginplot Matrix, Matrix Plot, Parallel Coordinates), resulting in nine experimental phases: *AM+Marginplot Matrix*, *AM+Matrix Plot*, *AM+Parallel Coordinates*, *JM+Marginplot Matrix*, *JM+Matrix Plot*, *JM+Parallel Coordinates*, *CM+Marginplot Matrix*, *CM+Matrix Plot* and *CM+Parallel Coordinates*. The study was designed as an online study and participants were recruited through E-mail lists for visualization and data analysis interest

groups. Each participant performed 36 tasks (4 per phase). None of the generated datasets were used more than once per participant, in order to minimize the risk that patterns that are asked for have been identified by chance during a previous tasks. The presentation order of the tasks and phases were counterbalanced using a Latin-square based procedure[30]. Equally many task were performed for each phase using the Iris and UKM based datasets. Dataset size was not treated as a factor in the experimental design, since all datasets used in the study were relatively small and thus no major performance difference was expected due to it. It was, however, recorded in order to be able to identify potentially unexpected results related to this. Performance was measured in terms of accuracy when performing the tasks. Response time was not considered a reliable enough measure in this particular context, since the study was conducted online rather than in a completely controlled environment, with an increased risk of participants being disrupted or taking a break during a task.

Procedure: The evaluation was carried out as an online survey implemented using the BOS Online Survey tool[31]. To ensure that participant had a basic level of understanding of missing data and missingness patterns, and the ability to interpret the visualization methods and tasks, the survey was initiated with a 13 minute introductory video explaining the details and concepts of relevance for the study. As an alternative, participants were provided the option of following a textual walk-through containing the same information as the video. 87% of participants choose to view the video. After the introduction a set of questions were asked to gather detail about the participants, their experience and perceived skill; including gender, age group, and experience of visualization methods, data analysis and missing data. This was followed by the main study where 36 tasks were presented, based on the questions described in the Task and Data section. The tasks were designed as multiple choice questions, presented along with a static image of the relevant visualization method displaying the appropriate dataset. The choice of using static images for the evaluation, rather than using an interactive environment, was made to reduce the impact of the varying interactive features, which would have needed to be measured carefully if included. Furthermore, the main aim was to measure the effectiveness of the visual attributes representing missing values, rather than the interactive features implemented in this particular tool. The answers of the participants were stored in BOS and later exported for statistical analysis of the results.

Results

23 participants finished the study, of which 5 female and 18 male. The biggest age group among the participants was 25-34 years (43.5%) followed by 35-44 years (21.7%), 45-54 years (17.4%), 55-64 years (13%) and 65 years or older (4.3%). Participants were asked to rank their level of experience of 1) visualization methods, 2) data analysis, and 3) missing data, using 5 point likert scales ranging from *No prior experience* (1) to *Professional* (5). A majority of participants (73.9%) ranked their experience of visualization methods high (4 or 5) while only 13% ranked their visualization experience low (1 or 2). Similar numbers were seen for data analysis experience, where 69.6% ranked themselves as highly experienced and only 8.6% as having low experience. When it came to experience of missing data the pattern was the opposite, with a small majority of 52.2% ranking their experience as low, while only 17.4% ranked their experience of missing data as high.

The remainder of this section will report on the results of the evaluation, in terms of accuracy when completing the tasks. To provide a comprehensive analysis of the results, taking the aspect of 'how wrong' an answer is into consideration, two sets of analysis were performed with different rankings of accuracy. In the first analysis an answer was considered either accurate or not accurate, only taking the number of accurately answered tasks into account. For the second analysis the answer was ranked from 0 to 3, depending on 'how wrong' the answer was; 3 was assigned to the correct answer, 2 two the wrong answer that was closest to the correct answer and so on. For example:

- For an AM question where the multiple choice answers to the question *Approximately how much data is missing in attribute D?* were *a)* 5%, *b)* 10%, *c)* 25% and *d)* 30%; and the correct answer was 25%, then *c* was ranked as 3, *d* was ranked as 2, *b* was ranked as 1 and *a* was ranked as 0.
- For a JM question where the multiple choice answers to the question *With which attribute does attribute B have the most jointly missing values?* were *a)* A, *b)* C, *c)* D and *d)* E, and the relative number of jointly missing values were *B&A*: 10%, *B&C*: 15%, *B&D*: 5% and *B&E*: 0%, then *b* was ranked as 3, *a* was ranked as 2, *c* was ranked as 1 and *d* was ranked as 0.
- For a CM question where the multiple choice answers to the question *Which trend in which attribute is most likely to be accountable for the missing data in attribute D?* were *a)* High in A, *b)* Low in B, *c)* High in C, and *d)* Low in C, and the CM relationships could be described as *D&A*: None, *D&B*: Low, *D&C*: Low (but more variation than in B), then *b* was ranked as 3, *d* was ranked as 2, *a* was ranked as 1 and *c* was ranked as 0.

While the main aim of the study was to investigate the performance of the combination of visualization methods and missingness patterns, the patterns and visualization were also analysed separately to find any overall patterns that may explain the detailed results.

Missingness Pattern

A Friedman test was used for significance testing, since the accuracy data was not normally distributed and could be described as ordinal or discrete rather than continuous. This was followed by a post-hoc test using a Wilcoxon signed-rank test and pairwise comparisons to identify for which missingness patterns the accuracy was significantly different. The descriptive statistics for accuracy for missingness patterns are presented in table 2 and figure 10, with results separated into *Accuracy*, where answers are considered accurate or not accurate and the maximum value was 12; and *Ranked Accuracy*, where answers are ranked based on how accurate they are and the maximum value was 36. The results showed an overall better performance when tasks related

Table 2. Descriptive statistics for the accuracy for missingness patterns

	Min	Median	Max
Accuracy			
AM	0	5	9
JM	3	7	12
CM	3	11	12
Ranked accuracy			
AM	10	23	31
JM	21	29	36
CM	18	34	36

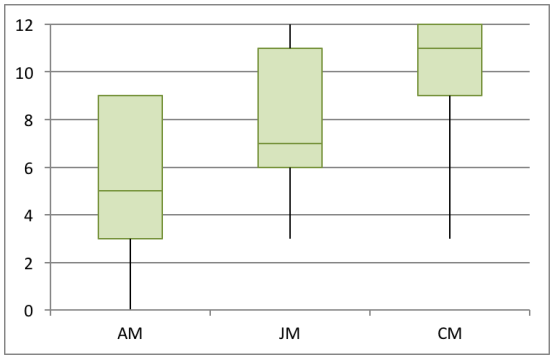
to *Conditional Missingness* (CM) was carried out in both analyses, while *Amount Missingness* (AM) tasks rendered the worst performance in both analyses. One potential reason why performance was better for CM may be that the visualization methods were originally designed for recorded data and, hence, may be more optimized for solving tasks related to recorded.

The statistical testing confirmed that the differences were significant, $\chi^2(2) = 32.116, p < 0.001$ for *Accuracy* and $\chi^2(2) = 16.636, p < 0.001$ for *Ranked Accuracy*. The post-hoc analysis with Wilcoxon signed-rank test was conducted with a Bonferroni correction applied, resulting in a significance level set at $p < 0.017$. The results displayed significant differences for all pairs of missingness patterns: $Z = -3.493, p < 0.001$ (*Accuracy*) and $Z = -3.636, p < 0.001$ (*Ranked Accuracy*) for AM vs JM; $Z = -4.032, p < 0.001$ (*Accuracy*) and $Z = -3.839, p < 0.001$ (*Ranked Accuracy*) for AM vs CM; and $Z = -3.801, p < 0.001$ (*Accuracy*) and $Z = -2.563, p = 0.01$ (*Ranked Accuracy*) for JM vs CM.

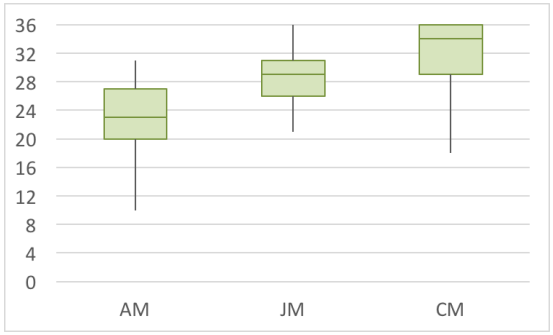
Visualization Method

The descriptive statistics for accuracy for visualization methods are presented in table 3 and figure 11, separated into analysis of *Accuracy* with maximum value 12, and analysis of *Ranked Accuracy* with maximum value 36. The results indicate better overall performance using Matrix Plot, compared to Marginplot Matrix and Parallel Coordinates, for both analyses; and worst overall performance using Marginplot Matrix.

Friedman tests confirmed significant differences for visualization methods, $\chi^2(2) = 26.847, p < 0.001$ for *Accuracy* and $\chi^2(2) = 21.816, p < 0.001$ for *Ranked Accuracy*. The post-hoc analysis with Wilcoxon signed-rank test was conducted with a Bonferroni correction applied, resulting in a significance level set at $p < 0.017$. The results displayed significant differences for all pairs of visualization methods for *Accuracy*: $Z = -3.967, p < 0.001$ for Marginplot Martix vs Matrix Plot; $Z = -3.750, p < 0.001$ for Marginplot Matrix vs Parallel Coordinates; and $Z = -2.646, p = 0.008$ for Matrix Plot vs Parallel Coordinates. For *Ranked Accuracy* the difference was significant for Marginplot Matrix vs Matrix Plot and Marginplot Matrix vs Parallel Coordinates ($Z = -3.531, p < 0.001$ and $Z = -3.536, p < 0.001$ respectively), while the difference for Matrix Plot vs Parallel Coordinates was not significant ($Z = -2.301, p = 0.021$). These results confirmed that overall, the Matrix Plot appear to perform better for identification of missingness patterns than the other two visualizations, while the Marginplot Matrix appear to perform worst.



(a) Accuracy



(b) Ranked accuracy

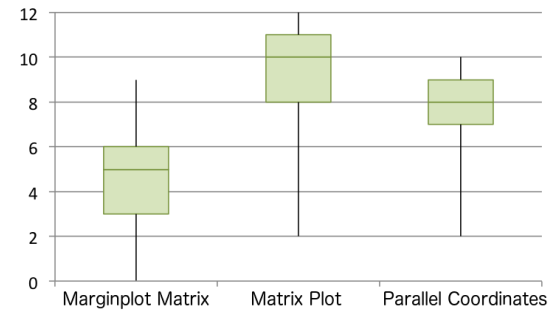
Figure 10. Distribution of accuracy results for the missingness patterns.

Table 3. Descriptive statistics for the accuracy for visualization vethods

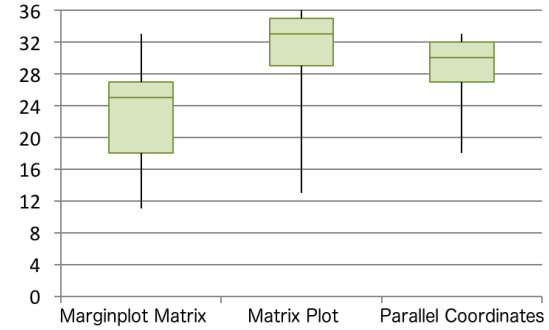
	Min	Median	Max
Accuracy			
Marginplot Matrix	0	5	9
Matrix Plot	2	10	12
Parallel Coordinates	2	8	10
Ranked accuracy			
Marginplot Matrix	11	25	33
Matrix Plot	13	33	36
Parallel Coordinates	18	30	33

Missingness Pattern and Visualization Method

The descriptive statistics for the nine phases of missingness patterns and visualization methods (table 4 and figures 12 and 13) show clear variations between different combinations, both for *Accuracy* where the maximum value is 4, and for *Ranked Accuracy* where the maximum value is 12. The combinations *JM+Matrix Plot*, *CM+Matrix Plot* and *CM+Parallel Coordinates* all display high accuracy with median values that are equal to the maximum values of



(a) Accuracy



(b) Ranked accuracy

Figure 11. Distribution of accuracy results for the visualization methods.

both analyses. The worst performances are in both analyses displayed for the combinations *AM+Marginplot Matrix*, *JM+Marginplot Matrix* and *AM+Parallel Coordinates*. It is worth noting that the worst performances for *Ranked Accuracy*, with median values of 7, is higher than 50% of the maximum value, while the worst performances for *Accuracy*, with median values of 1, is considerably worse landing at 25% of the maximum value. This may indicate that while responses were wrong, they were not necessarily very far from the correct answer.

Statistical testing with Friedman tests confirmed significant differences for the interaction of visualization method and missingness patterns, $\chi^2(8) = 103.776, p < 0.001$ for *Accuracy* and $\chi^2(8) = 90.523, p < 0.001$ for *Ranked Accuracy*. This was followed by post-hoc analysis with Wilcoxon signed-rank test and pairwise comparisons to identify for which combinations of visualization method and missingness pattern the accuracy was significantly different, applying a Bonferroni correction resulting in a significant level set at $p < 0.0056$. The results displayed significant differences for some combinations of visualization methods and missingness patterns.

When it comes to drawing meaningful conclusions based on performance differences for combinations of visualization methods and missingness patterns, some pairs of combinations are more relevant than others. These are the pairs were either the missingness pattern or the visualization method is the same. The following interesting results were found through the post-hoc analysis for *Accuracy*.

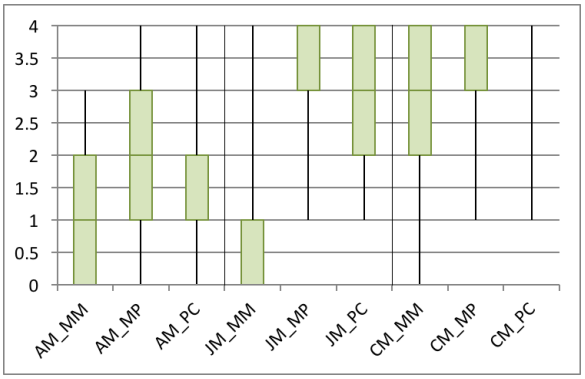
Amount Missing: Examining the performance of the different visualization methods for AM tasks, significant differences were found for Marginplot Matrix vs Matrix Plot ($Z = -3.156, p = 0.002$, first and second box-plot in figure 12a) and for Matrix Plot vs Parallel Coordinates ($Z = -2.918, p = 0.004$, second and third box-plot in figure 12a). This indicates that Matrix Plot performs better than both Marginplot Matrix and Parallel Coordinates for tasks related to AM patterns. While the accuracy results indicate that Parallel Coordinates may have performed better than Marginplot Matrix, the difference was not significant.

Joint Missingness: For tasks relating to JM patterns, significant differences were found for all visualization methods (Marginplot Matrix vs Matrix Plot: $Z = -3.829, p < 0.001$; Marginplot Matrix vs Parallel Coordinates: $Z = -3.472, p = 0.001$, and Matrix Plot vs Parallel Coordinates: $Z = -2.944, p = 0.003$ respectively). The performance was better for Matrix Plot than Marginplot Matrix (fifth and fourth box-plot in figure 12a), and for Parallel Coordinates than for Marginplot Matrix (sixth and fourth box-plot in figure 12a). The Matrix Plot also performed better than Parallel Coordinates (fifth and sixth box-plot in figure 12a).

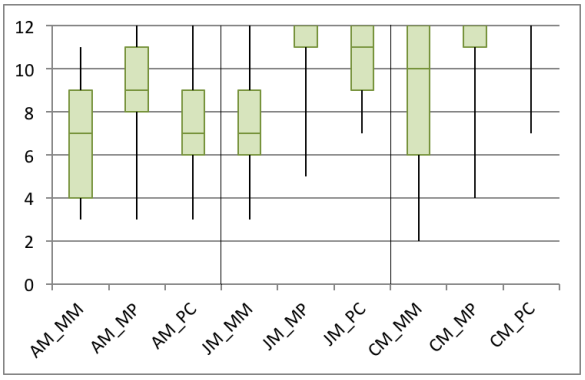
Conditional Missingness: There were no significant performance differences for the visualization methods for CM tasks, however the combination Marginplot Matrix vs Parallel Coordinates (seventh and ninth box-plot in figure 12a) has a relatively low p-value ($Z = -2.729, p = 0.006$) indicating that Parallel Coordinates likely performs better than Marginplot Matrix for tasks related to CM patterns. The accuracy result for Parallel Coordinates is also slightly better than for Matrix Plot, but the difference is too small to draw any conclusions from it.

Table 4. Descriptive statistics for the accuracy for the nine phases

	Min	Median	Max
Accuracy			
AM+Marginplot Matrix	0	1	3
AM+Matrix Plot	0	2	4
AM+Parallel Coordinates	0	1	4
JM+Marginplot Matrix	0	1	4
JM+Matrix Plot	1	4	4
JM+Parallel Coordinates	1	3	4
CM+Marginplot Matrix	0	3	4
CM+Matrix Plot	1	4	4
CM+Parallel Coordinates	1	4	4
Ranked accuracy			
AM+Marginplot Matrix	3	7	11
AM+Matrix Plot	3	9	12
AM+Parallel Coordinates	3	7	12
JM+Marginplot Matrix	3	7	12
JM+Matrix Plot	5	12	12
JM+Parallel Coorindates	7	11	12
CM+Marginplot Matrix	2	10	12
CM+Matrix Plot	4	12	12
CM+Parallel Coordinates	7	12	12

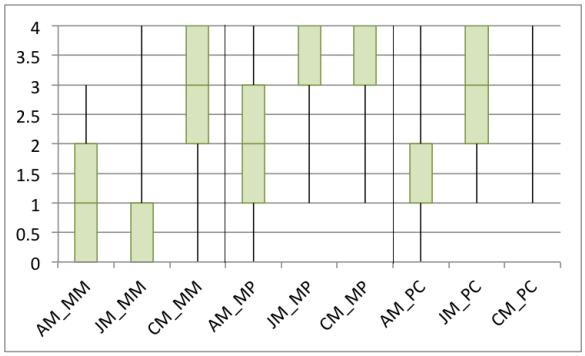


(a) Accuracy

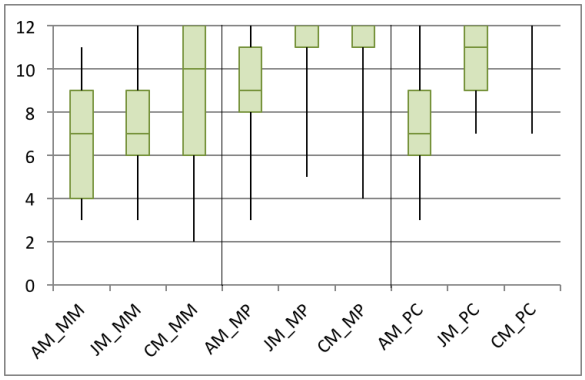


(b) Ranked accuracy

Figure 12. Distribution of accuracy results for the combinations of missingness patterns and visualization method, ordered by pattern. MM = Marginplot Matrix, MP = Matrix Plot and PC = Parallel Coordinates.



(a) Accuracy



(b) Ranked accuracy

Figure 13. Distribution of accuracy results for the combinations of visualization method and missingness patterns, ordered by visualization. MM = Marginplot Matrix, MP = Matrix Plot and PC = Parallel Coordinates.

Marginplot Matrix: Examining the accuracy results for the Marginplot Matrix (left section in figure 13a) there was a significant difference for AM vs CM patterns and for JM vs CM patterns, $Z = -3.311, p = 0.001$ and $Z = -3.209, p = 0.001$ respectively, with better performance for CM tasks than for AM and JM tasks when using the Marginplot Matrix (as visible from the third, first and second box-plots respectively in figure 13a).

Matrix Plot: The Matrix Plot (middle section in figure 13a) shows significantly better performance for JM tasks than for AM tasks with $Z = -3.678, p < 0.001$ (fifth and fourth box-plots in figure 13a), as well as significantly better performance for CM tasks than for AM tasks with $Z = -3.678, p < 0.001$ (sixth and fourth box-plots in figure 13a), which agrees with the overall result with generally worse performance for AM patterns.

Parallel Coordinates: When it comes to Parallel Coordinates, all differences for missingness patterns were significant, with higher performance for JM than AM (eight and seventh box-plots in figure 13a, $Z = -3.313, p = 0.001$), higher performance for CM than AM (ninth and seventh box-plots in figure 13a, $Z = -4.149, p < 0.001$) and higher performance for CM than JM (ninth and eight box-plots in figure 13a, $Z = -3.640, p < 0.001$).

The post-hoc analysis on *Ranked Accuracy* displayed similar results, however a few of the interesting differences were no longer significant. The performance difference between AM and CM patterns for the Martginplot Matrix (first and third box-plots in figure 13b) was not significant. Similarly, the previously significant difference between Matrix Plot and Parallel Coordinates for JM patterns (fifth and sixth box-plots in figure 12b) was no longer significant.

Discussion

As demonstrated in the pilot study, there are some significant performance differences between methods for visualizing missing data that depend on the examined missingness pattern. This section will discuss the results and their indications in terms of guidance for future design of missing data visualization.

The Matrix Plot performed best overall, and also best for tasks related to *Amount Missing* (AM) and *Joint Missingness* (JM) patterns, with statistically significant differences. The Matrix Plot performed significantly better for JM tasks than for AM tasks, and also for *Conditional Missingness* (CM) task in comparison to AM, while displaying identical performance for JM and CM tasks. This could possibly indicate that the Matrix Plot is better for identifying CM and JM patterns than it is for identifying AM patterns. It is however worth keeping in mind that tasks relating to CM and JM patterns had overall significantly better performance than tasks relating to AM patterns. This may indicate that AM patterns are more difficult to perceive overall, for all three visualization methods, suggesting that this type of pattern may need further attention when designing visualization methods, particularly in context of visualization methods

for numerical data where the frequency of a certain value may often be less clearly visualized than in methods for categorical data. Linking to the visual attributes in table 1 the Matrix Plot represents missing values by colour and not location, the visual representation partly represent frequency through size by separating items with same values, and in part connecting across variables and connecting missing and recorded values. Furthermore, the Matrix Plot does not separate missing and recorded values in the display. A conclusion that may be drawn based on this is that the colour combined with frequency/size representation had a positive impact on frequency related tasks (AM and JM), while the in-part connection across variables and between missing and recorded, as well as the mixed representation of missing and recorded may have had a positive impact on CM tasks in comparison to the Marginplot Matrix, which does not have these features. With the ability for supporting identification of, in particular, frequency related patterns, the Matrix Plot may be useful for indication or exclusion of the MCAR missingness mechanism. It may also indicate subsets of data where the missingness is particularly problematic due to the frequency. Similarly, it may be useful for indicating potential MNAR situations through its ability to support identification of JM patterns. While the Matrix Plot overall performed well, it may benefit from enhanced representation of CM patterns with clearer links between recorded and missing values. Generally, colour may not be as reliable for representation of numerical values or differences as, for instance, location or size.

Parallel Coordinates had the second best performance overall, with overall performance closer to Matrix Plot than to Marginplot Matrix. Parallel Coordinates performed better than Matrix Plot and Marginplot Matrix for CM tasks, although the differences were not statistically significant. There was a significant performance difference (in favour for Parallel Coordinates) for JM patterns in comparison to Marginplot Matrix. The visual attributes used by Parallel Coordinates to represent missingness includes a selectable in-part use of colour (highlighting missing values of a selected variable), using a separate location for missing values, and items with the same value being drawn on top of each other resulting in there being no direct frequency/size representation. Parallel Coordinates has clear connection across variables and between missing and recorded values, through this only partly separates missing and recorded. The results of the study indicate that the clear connection features of Parallel Coordinates may have had a positive impact on CM tasks and it may, hence, be useful particularly for identification of the MAR missingness mechanism (and exclusion of MCAR) and the hiding concept[27], as well as for supporting selection of imputation method, as strong CM patterns may suggest imputation based on a subset of the recorded data. The connectedness may also have had a positive impact on JM results in comparison to Marginplot Matrix, which overall emphasize the importance of connectedness for identification of multivariate patterns. Furthermore, the colour use in Parallel Coordinates may potentially be of higher importance for distinguishing between missing and recorded values than the colour use in Marginplot Matrix, due to the way the Parallel Coordinate colouring supports identification of patterns that link missingness patterns across multiple variables. The main limitation of Parallel Coordinates, in context of identification of missingness patterns, is its inability to clearly display frequencies of missing values, which is also well known from visualization of categorical values in Parallel Coordinates. This needs to be taken into consideration for future designs with visual enhancements representing frequency, as for instance briefly suggested by Fernstad and Glen[3]. Another issue that is not measured in this study, but discussed earlier in the paper, is the risk of misinterpreting the missing values as high values when they are located above the axis, which could be reduced by more clearly indicating that the missing values are indeed different from the recorded values.

The Marginplot Matrix had the worst overall performance of the visualization methods, and also performed worst when breaking down the results for the different missingness patterns (although all differences were not significant). In particular, Marginplot Matrix performed worse for tasks relating to JM patterns, where the differences to both Matrix Plot and Parallel Coordinates were significant; as well it performed worse for identification of CM patterns. Nonetheless, the performance for CM patterns using Marginplot Matrix was significantly better than the performance for both AM and JM patterns using Marginplot Matrix. It is nteresting to note that the Marginplot Matrix performed slightly better for AM tasks than it did for JM tasks when looking at accuracy, while the overall performance of all visualization methods is better for JM than AM. The difference between AM and JM is however not significant for Marginplot Matrix. Linking to the visual attributes, as shown in table 1, the Marginplot Matrix represent missing values through location where missing values are separated from recorded values. Colour is also used to emphasize the difference but it is not the main visual attribute. Items with the same values in two variables are drawn on top while items with the same value in only one variable are not. This results in a part representation of frequency through size, but only for items that have recorded values for one variable. This may be one reason why Marginplot Matrix performed worse for JM than AM, as AM only addresses missingness in one variable while only items with missing values in both variables are relevant for the JM pattern. Furthermore, the lack of connectedness across variables and between missing and recorded may have impacted negatively on the JM and CM tasks. Marginplot Matrix use the size of a box-plot to represent the distribution of recorded values in one variable that are missing in the other variable in comparison to the values that are recorded in the other variable. While an initial hypothesis was that this would have positive impact on the detection of CM patterns, it does not seem to have been as useful as the attributes used in Parallel Coordinates and Matrix Plot. Similarly to Parallel Coordinates, a main issue with the Marginplot Matrix in the context of missingness patterns is its limited ability to display frequency patterns, and in particular in terms of joint missingness across multiple variables. This could possibly be improved through the inclusion of size-based visual representation of missingness frequency, as also suggested for Parallel Coordinates. Another issue which was not examined in great detail in this study is the general scalability issue of small multiple displays, and the difficulty

of perceiving detail as the number of variables in the dataset increase. For future designs, building on scatterplot matrix approaches, the trade-off between representing more missingness detail (such as additional visual enhancement for frequency representation) and the amount of detail that can be efficiently displayed using small multiples, has to be carefully considered. Furthermore, the pilot study presented in this paper indicates that Matrix Plot and Parallel Coordinates may be more useful for identifying missingness patterns, hence suggesting that it may be more beneficial to build upon these approaches rather than the Marginplot Matrix.

Overall, the evaluation results emphasize the importance of visually representing missingness frequency when it comes to identifying AM and JM patterns, and to carefully consider the ability of a visualization design to display the relevant patterns for missingness analysis. To summarize, the results of the study indicates the importance of clear frequency representation as well as the importance of connectedness across variables and between missing values and recorded values. It furthermore seems beneficial not to separate missing and recorded values, especially in the context of identifying CM patterns, but there is nothing in the results that indicate that it would be negative for identification of any patterns. While certain aspects and visual attributes may be of more or less relevance depending on data and application area, it is likely that all three missingness patterns are of relevance for understanding the missingness in data. While the study presented in this paper is merely a first step in identifying appropriate visual enhancements for representation of missing values and missingness patterns, some conclusions can be drawn to guide and support future research in missing data visualization. It is likely to be beneficial for the analysis of missingness in data to a) include size based representation of the frequency of missing values, b) utilize visual features that connect missing and recorded values across multiple variables, and c) to avoid separating the missing and recorded values into two sets of representations. Furthermore, based on prior research and the results from this study, the author would suggest that location is not suitable as the only representation of missing values, due to the risk of misinterpretation in visualization where location has a meaning (in part relating to the misleading display as evauated by Eaton et al.[26]). When such representations are used, some additional visual feature, such as colour, should preferably be used to emphasize the missingness of values.

Conclusions and Future Work

Even though missing data is commonly occurring in almost every data generating domain, very little effort has been put into the visualization of missing values, as highlighted in this paper. The understanding of missing values and the patterns underlying the missingness are nonetheless important both for understanding the cause of the missingness and how to best deal with it, as well as to gain a more extensive understanding of the data as a whole. While in a unique position to facilitate missingness data analysis, few attempts have been made to design visualization tools for supporting visual investigation of missingness in data.

This paper presented and motivated a set of missingness patterns of relevance for investigation and understanding of missing values in data. These patterns include *Amount Missing*, which represents the relative number of items with missing values in a variable or item; *Joint Missingness*, which refers to the quantity of items that have missing values for both variables in a variable pair; and *Conditional Missingness*, which refers to relationships between missing values in one variable and recorded values in another.

The paper also contributed an initial usability evaluation where visualization methods that use different visual attributes for representing missing values are compared in context of identification of the three missingness patterns. The evaluation results indicate that a Matrix Plot, a heatmap where missing data is represented by colour, is generally the best of the visualization methods when it comes to performing tasks related to *Amount Missing* and *Joint Missingness*. This may be due to its more straightforward ability to display frequency compared to the other visualization methods. For tasks related to *Conditional Missingness* the Parallel Coordinates, where missing values are represented above the axis and items with missing values in a selected variable are highlighted, performed slightly better than the Matrix Plot. This may indicate that, in context of understanding relationships between missing and recorded values, the lines connecting values across multiple variables, as in Parallel Coordinates, may be of more benefit than the rows and colour value representation of the Matrix Plot. The third visualization method, Marginplot Matrix, performed worse than both Parallel Coordinates and Matrix Plot for all missingness patterns. Explanations for this may possibly be a consequence of that the Marginplot Matrix represent missing values in the margin, separated from the recorded values, and plots items with the same value on top of each other, hence affecting the ability to perceive frequencies. The small size of plots when using small multiples, as in the Marginplot Matrix, may also have had a generally negative impact on pattern identification.

While the work presented in this paper provide some initial guidance on visual attributes to use when designing visualization methods for missing data, further research is needed to establish which methods may best support understanding of missingness in data. This includes studies on more complex and larger data, as well as further studies of visual attributes. Further research also includes the development of visual analytics tools that are able to support interactive exploration of missingness in large and heterogeneous datasets and that support decision making in terms of how best to deal with the data.

Acknowledgements

The author would like to thank Dr Matthias Templ at Vienna University of Technology, Austria, for his input on the study and missingness patterns, and to James Fraser, student at Northumbria University, UK, for conducting a pre-pilot study on the subject. Thanks also to scientists at Unilever R&D Port Sunlight, UK, in particular Dr Tim Madden, Dr Stephen Bennett and Dr Jane Shaw, for input and feedback on practical aspects of missing data analysis.

References

1. Kandel S, Heer J, Plaisant C et al. Research directions in data wrangling: Visualizations and transformations for usable and credible data. *Information Visualization* 2011; 10(4): 271–288.
2. Wong BLW and Varga M. Black holes, keyholes and brown worms: Challenges in sense making. In *Proceedings of the Human Factors and Ergonomics Society 56th Annual Meeting*. pp. 287–291.
3. Johansson Fernstad S and Glen RC. Visual analysis of missing data – to see what isn't there. In *Poster Proceedings of IEEE Vis*. IEEE.
4. Kirk A. Visualizing zero: How to show something with nothing. http://blogs.hbr.org/2014/05/visualizing-zero-how-to-show-something-with-nothing/, 2014.
5. Rubin DB. Inference and missing data. *Biometrika* 1976; 63(3): 581–592.
6. Bache K and Lichman M. UCI machine learning repository, 2013. URL http://archive.ics.uci.edu/ml.
7. Fielding S, Fayers PM and Ramsay CR. Investigating the missing data mechanism in quality of life outcomes: a comparison of approaches. *Health and Quality of Life Outcomes* 2009; 7(1): 57.
8. Djurcilov S and Pang A. Visualizing sparse gridded data sets. *IEEE Computer Graphics and Applications* 2000; 20(5): 52–57.
9. Twiddy R, Cavallo J and Shiri SM. Restorer: A visualization technique for handling missing data. In *Proceedings of the conference on Visualization'94*. IEEE Computer Society Press, pp. 212–216.
10. Unwin A, Hawkins G, Hofmann H et al. Interactive graphics for data sets with missing values manet. *Journal of Computational and Graphical Statistics* 1996; 5(2): 113–122.
11. Theus M, Hofmann H, Siegl B et al. Manet extensions to interactive statistical graphics for missing values. In *In New Techniques and Technologies for Statistics II*. IOS Press, pp. 247–259.
12. Templ M, Alfons A and Filzmoser P. Exploring incomplete data using visualization techniques. *Advances in Data Analysis and Classification* 2012; 6(1): 29–47.
13. Brix P. miP: Multiple imputation plots. http://ftp.cs.pu.edu.tw/network/CRAN/web/packages/miP /miP.pdf, 2011.
14. Schulz HJ, Nocke T, Heitzler M et al. A systematic view on data descriptors for the visual analysis of tabular data. *Information Visualization* 2017; 16(3): 232–256.
15. Carpenter B, Goodrich B and Su YS. migui: Graphical user interface to the 'mi' package. https://cran.r-project.org/web/packages/migui/migui.pdf, 2015.
16. Su YS, Gelman A, Hill J et al. Multiple imputation with diagnostics (mi) in r: Opening windows into the black box. *Journal of Statistical Software* 2011; 45(2): 1–31.
17. Honaker J, King G and Blackwell M. Amelia ii: A program for missing data. *Journal of Statistical Software, Articles* 2011; 45(7): 1–47.
18. Swayne DF and Buja A. Missing data in interactive high-dimensional data visualization. *Computational Statistics* 1998; 13(1): 15–26.
19. Swayne DF, Lang DT, Buja A et al. Ggobi: Evolving from xgobi into an extensible framework for interactive data visualization. *Comput Stat Data Anal* 2003; 43(4): 423–444.
20. Cheng X, Cook D, Hofmann H et al. Visually exploring missing values in multivariable data using a graphical user interface. *Journal of Statistical Software* 2015; 68(6): 1–23.
21. Cedilnik A and Rheingans P. Procedural annotation of uncertain information. In *Visualization 2000. Proceedings*. IEEE, pp. 77–84.
22. Xie Z, Huang S, Ward MO et al. Exploratory visualization of multivariate data with variable quality. In *In Proceedings of the IEEE Symposium on Visual Analytics Science and Technology*. pp. 183–190.
23. Arbesser C, Spechtenhauser F, Mühlbacher T et al. Visplause: Visual data quality assessment of many time series using plausibility checks. *IEEE Transactions on Visualization and Computer Graphics* 2017; 23(1): 641–650.
24. Wang H and Wang S. Visualization of the critical patterns of missing values in classification data. In *Advances in Visual Information Systems*. Springer, 2007. pp. 267–274.
25. Kohonen T. The self-organizing map. *Neurocomputing* 1998; 21(1–3): 1–6.
26. Eaton C, Plaisant C and Drizd T. Visualizing missing data: graph interpretation user study. In *Human-Computer Interaction-INTERACT 2005*. Springer, 2005. pp. 861–872.
27. Wang H and Wang S. Data mining with incomplete data. In *Encyclopedia of Data Warehousing and Mining*, second ed. IGI Global, 2009. pp. 526–530.
28. Templ M and Filzmoser P. Visualization of missing values using the R-package VIM. *Reserach report cs-2008-1, Department of Statistics and Probability Therory, Vienna University of Technology* 2008; .
29. Thompson J, Blair M, Chen L et al. Video game telemetry as a critical tool in the study of complex skill learning. *PLoS ONE* 2013; 8(9). DOI:10.1371/journal.pone.0075129.
30. Graziano AM and Raulin ML. *Research methods: A process of inquiry* . HarperCollins College Publishers, 1993.
31. of Bristol U. Bos online survey tool. https://www.onlinesurveys.ac.uk/, 2016.